

# PSI – BLAST

## Position Specific Iterative

### BLAST

Encadrants : M. Frédéric Dardel  
M. Nicolas Loménie

Rédigé et présenté par : MABROUK Skandar

# Table des matières

- 1) Définition
- 2) Principe général
- 3) Construction de PSSM
- 4) PSI-BLAST vs BLAST
- 5) Pourquoi le programme nécessite-t-il plusieurs cycles d'itération ?  
Qu'est qui change à chaque cycle ?
- 6) avantages et inconvénients
- 7) Quelle information peut-on tirer du profil de score construit par  
PSI-BLAST et quelles utilisations peut-on en faire ?
- 8) Test et comparaison



# Définition

- PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST.

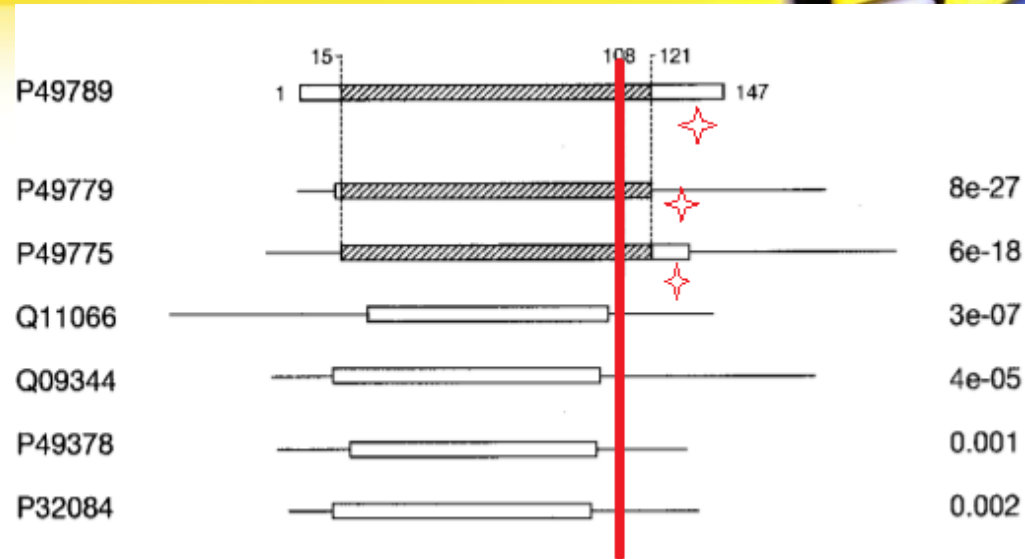
# Principe Général

## Les étapes principales de l'algorithme consiste à :

- I. effectuer une recherche BLAST standard en utilisant une matrice de substitution BLOSUM 62 avec une seule séquence d'interrogation.
- II. obtenir un ensemble initial de séquences apparentées dont le score BLAST donne une E-value inférieure à un seuil prédéterminé. (« On doit faire attention à la sélection de séquence »)
- III. PSI-BLAST construit un alignement de séquences multiples puis crée un "profil" ou une matrice de score spécifique à la position (PSSM).
- IV. Comparez la PSSM à la base de données en utilisant une variante du programme BLAST pour identifier de nouvelles séquences avec des E-values suffisamment petites (estimer la signification) .
- V. Répétez l'étape IV. De manière itérative : à chaque nouvelle recherche un nouveau profil/PSSM est utilisé comme requête.

# Construction de PSSM

- A cette position nous avons uniquement 3 résidus qui seront considérés .
- Le PSSM est limité aux résidus qui ont été alignés sur un résidu de la séquence d'interrogation.
- A chaque position de résidu de la requête, un PSSM est construit en utilisant seulement les séquences dont les alignements BLAST impliquent cette position.
- Le nombre de séquences dans l'alignement change d'une colonne à l'autre.



# PSSM



<b>Avantages PSSM</b>	<b>Inconvénients PSSM</b>
Bonne méthode pour de courtes régions conservées.	Insertions et délétions interdites avec les PSSM, sinon utiliser les profils généralisés .
Approche statistique (basée sur la taille des banques) (E-Value)	Les séquences correspondant à de longues régions ne peuvent être décrites avec cette méthode.



# PSI-BLAST contre BLAST



- En raison de sa nature cyclique, PSI BLAST permet de trouver des homologues plus éloignés qu'une simple recherche BLAST.
- PSI BLAST utilise deux valeurs E :
  - « threshold » Le seuil E-valeur pour le BLAST initial (10 par défaut).
  - « inclusion » Les E-valeurs d'inclusion pour accepter la séquence dans la construction du PSSM (0.001 par défaut).

Pourquoi le programme nécessite-t-il plusieurs cycles d'itération ? Qu'est qui change à chaque cycle ?

=> Avec les itérations, le PSSM continue d'être mis à jour, ce qui rend le BLAST plus sensible à la recherche d'homologues.

=> Après chaque itération :

- Génération d'un nouveau alignement multiple en rassemblant les alignement dont la E-value est inférieure à un seuil défini.



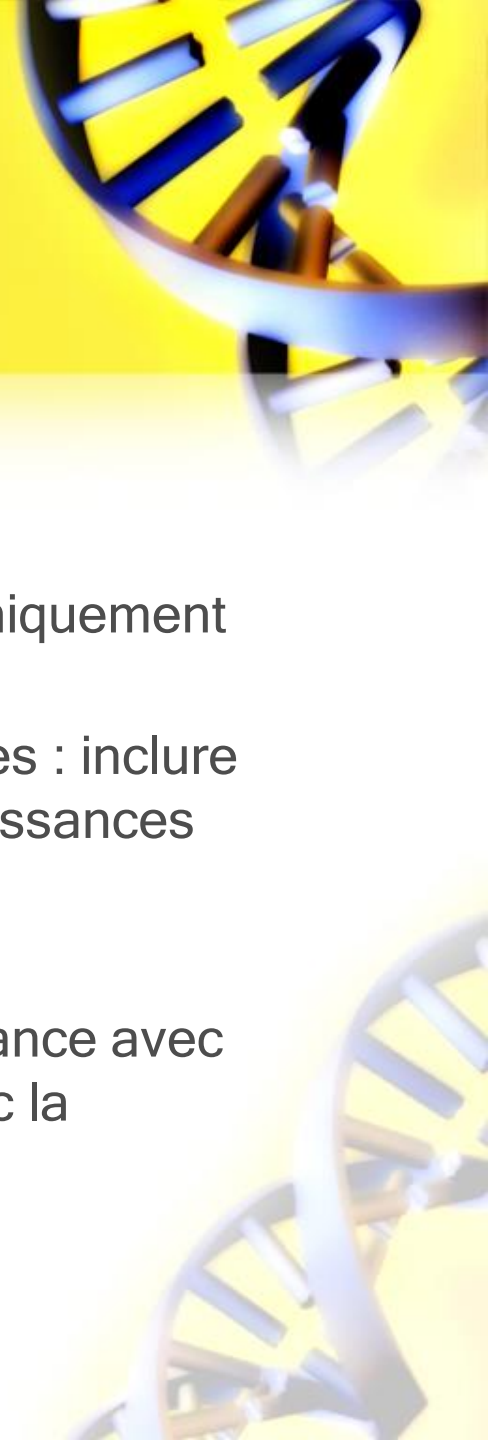
# PSI BLAST AVANTAGES



- RAPIDE GRÂCE À L'HEURISTIQUE BLAST
- Recherche de PSSM sur de grandes bases de données
- Un Traitement statistique très sophistiqués de score de correspondance.
- Un algorithme particulièrement efficace pour la pondération des séquences :
  - LE SCHÉMA DE PONDÉRATION DES SÉQUENCES BASÉ SUR LA POSITON, LÉGÈREMENT MODIFIÉ POUR INCLURE LES LACUNES COMME UN AUTRE TYPE DE RÉSIDU, ET POUR IGNORER LES RÉSIDUS ENTIÈREMENT CONSERVÉS.

# Inconvénients

- Eviter les séquences trop proches => OVERFIT !
- Peut inclure des faux homologues ! Mais c'est uniquement avec des séquences pas bien faites.
- Il faut vérifier soigneusement les correspondances : inclure ou exclure les séquences en fonction des connaissances biologiques .
- La E-value reflète l'importance de la correspondance avec l'ensemble d'entraînement précédent et non avec la séquence originale




# La corruption des profils

- Les E-valeurs de PSI-BLAST sont calculées pour les profils produits par PSI-Blast, et ne peuvent pas être interprétées comme une référence à la séquence originale de la requête.
- Une fois qu'une séquence sans rapport avec la requête est incluse dans un alignement multiple PSI-BLAST, donc dans la construction du profil de PSI-BLAST, elle amènera beaucoup de ses "voisins" à l'itération suivante, et ce processus peut faire un effet de boule de neige : La pondération des séquences va exacerber ce processus.

# TEST ON A HUMAN PROTEIN



>sp|P02144|MYG\_HUMAN Myoglobin OS=Homo sapiens GN=MB PE=1 SV=2  
MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGHPETLEKFDKFKHLKSEDEMKASE  
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH  
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

- Après 3 itérations , on n'en retrouve plus de nouveau homologues de la myoglobine :
  - En changeant le threshold cela peut réduire le nombre d'itérations nécessaire pour obtenir le PSSM
- 

Pendant la première itération nous retrouvons le résultat de blast normal ensuite nous retrouvons les nouvelles séquences ,  
à la deuxième et la troisième itération => les séquences deviennent « old » et leur E-value changeant .

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	SP:P02144	Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=1 SV=2	154	316.0	100.0	100.0	4.0E-112
		Cross-references and related information in: ► Gene expression ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Molecular interactions ► Protein families ► Diseases ► Macromolecular structures ► Protein expression data ► Reactions & pathways ► Protein sequences					
2	TR:A0A1K0FU49	Myoglobin OS=Homo sapiens OX=9606 GN=GLNG PE=3 SV=1	154	316.0	100.0	100.0	4.0E-112
		Cross-references and related information in: ► Gene expression ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein families ► Protein sequences					
3	TR:A0A024R1G3	Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=3 SV=1	166	316.0	100.0	100.0	7.0E-112
		Cross-references and related information in: ► Nucleotide sequences ► Literature ► Samples & ontologies ► Protein families ► Protein sequences					
4	TR:B2RA67	Myoglobin OS=Homo sapiens OX=9606 PE=2 SV=1	154	311.0	99.0	99.0	2.0E-110

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	TR:B2RA67	Myoglobin OS=Homo sapiens OX=9606 PE=2 SV=1	154	208.0	99.0	99.0	3.0E-69
		Cross-references and related information in: ► Nucleotide sequences ► Samples & ontologies ► Protein families ► Protein expression data ► Protein sequences					
2	SP:P02144	Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=1 SV=2	154	207.0	100.0	100.0	6.0E-69
		Cross-references and related information in: ► Gene expression ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Molecular interactions ► Protein families ► Diseases ► Macromolecular structures ► Protein expression data ► Reactions & pathways ► Protein sequences					
3	TR:A0A1K0FU49	Myoglobin OS=Homo sapiens OX=9606 GN=GLNG PE=3 SV=1	154	207.0	100.0	100.0	6.0E-69
		Cross-references and related information in: ► Gene expression ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein families ► Protein sequences					
4	TR:A0A024R1G3	Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=3 SV=1	166	207.0	100.0	100.0	9.0E-69
		Cross-references and related information in: ► Nucleotide sequences ► Literature ► Samples & ontologies ► Protein families ► Protein sequences					



## Les acides aminés jusqu'à la fin de la séquence

Taille L\*20  
avec  
L=length(seq)

Un même acide aminé peut prendre différent score en fonction de sa position

Aux positions suivantes pour chaque acide aminé représenté on visualise un énorme <trou> est ceci est expliqué par la non-similarité

Une telle fréquence peut entraîner l'exclusion de ces acides aminés à ces positions

on dit des substitutions extrêmement défavorables !

	K	Lambda
Standard Ungapped	0.1353	0.3161
Standard Gapped	0.0374	0.2670
PSI Ungapped	0.1223	0.3170
PSI Gapped	0.0374	0.2670



>SP:P68871 HBB\_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB  
PE=1 SV=2  
Length=147

Score = 197 bits (501), Expect = 3e-65  
Identities = 36/145 (25%), Positives = 58/145 (40%), Gaps = 2/145 (1%)

début

```
Query 3 LSDGEWQLVLMVWGKV EADIPGHGQEV LIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL 62
      L+ E V +WGKV ++ G E L RL +P T F+ F L + D + + +
Sbjct 4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 61

Query 63 KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG 122
      K HG VL A L + + L++ H K + + + ++ VL
Sbjct 62 KAHGKKVLGAFSDGLAHLNLTGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 121

Query 123 DFGADAQGAMNKAL ELFRKDIASNY 147 fin
      +F Q A K + + A Y
Sbjct 122 EFTPPVQAAYQKVVAGVANALAHKY 146
```

Il y a un 25% de similarité entre les deux séquences.

En bleu : les séquences sont homologues

En vert : il y a un GAP, pas de similarité à cette position .

**Thanks  
For  
Watching !**

